State of California
Office of the State Chief Information Officer

# Enterprise Architecture Standard

**Geocode Standard**
**Reference Model Type and ID No: SRM 74.742.590.1**
**Status:** Proposed
**Analysis:** OCIO Geographic Information Systems and Enterprise Architecture
**Effective Date**: mm/dd/yyyy
**Next Review Date**: mm/dd/yyyy
**Approved By:** Office of the State Chief Information Officer (OCIO)

## Introduction

Geographic data exists in State databases and files that describe locations such as street addresses, city names, ZIP Codes, and even telephone numbers. While people understand what these descriptions mean and how they relate to locations on the earth's surface, analysis for the intent of decision making in a computer can do little with this information. Geocoding is the process of assigning "X, Y" coordinate value to a place by comparing the descriptive location elements to those present in a reference data set. These X, Y coordinate pairs are point locations which can be displayed on a digital map and prepared for use in analysis.

Once the location information is transformed into a common two-dimensional coordinate system (X and Y) various analytical techniques can be performed that can disclose patterns and relationships that are not apparent in tabular data. Patterns that can be analyzed include, but are not limited to;

| Issue | Description |
|---|---|
| Fraud | Potential fraud cases because the normal distribution of government funds (e.g. unemployment insurance or public health assistance) is far exceeded at a given address |
| Health Patterns | Identification of health concerns, like clusters of a disease, outside the normal rate of infection and hospitalization issues from a set of addresses within a neighborhood |
| Environmental Patterns | Evaluation of environmental concerns given the concentration of toxic release issues within a range of addresses |
| Taxation | Assessment of fair tax collection and levees given the normal distribution of taxation at a set of addresses |
| Economic Development | Strategic planning of business development and growth in the state given addresses of employer locations and access to business tax incentives |
| Government Services | Identification for the need for government services given the types of populations at a set of addresses |
| Government Services | Evaluation of the effectiveness, equity and transparency of government services at a set of addresses |

Data that is geocoded has a common element for database management, the X, Y coordinates. These common elements can be used to help populate and effectively manage ancillary data often required

for the state's data systems, because the X, Y is a unique identifier for place. This unique identifier can be used to populate the ancillary data. This ancillary data can be elements like the county name, legislative district, and special district. Moreover, geocoding can standardize the descriptive data so that there is a single collection of real identified addresses, street names, cities and ZIP codes. If proper geocoding techniques are used, fields like these can be automatically populated reducing human entry time and error caused by human data entry. A proper geocoding infrastructure can help more effectively manage our State databases.

Geocoding has three main components: 1) Reference Data, 2) Address (or source) Data, and 3) Software (or geocoding applications):

### 1. Reference Data

Reference data is the geographic base files on which a geocoding engine runs. These base files are most often address point locations, street center line data, ZIP Code boundaries and place name locations. Reference data is available from both public and private sources. Public data sources include the US Census Bureau TIGER Line files, and the US Postal Service city/state five digit ZIP codes and ZIP+4 files. Numerous private sources provide licenses to reference data depending on business need.

The reference data must be stored and maintained in the standard projection and datum. Establishing this allows for the geocoding service to return the standardized projection for the X, Y coordinates. A projection algorithm can be used to return the Latitude and Longitude of the X, Y in the standard datum [See Geocode Projection Standard, Statewide Information Management Manual (SIMM) Section 58D].

### 2. Address (or Source) Data

Address or source data is the descriptive place data owned by the business application. These data vary by type but, in terms of State government, it is most often a physical street address (example: 1325 J Street). In addition, nearly every department in State government has a data set of street address locations for managing its own building location services and outlets for public access. Source data can also be incident or event locations (e.g. a fire, or vehicle accident), locations of equipment and facilities (e.g. medical stockpile locations, hospitals), and monument locations (e.g. .1 miles north from intersection of Hwy 49 and Interstate 80 on Hwy 49).

**Note:** It is a State of California practice to recognize that a street address geocode constitutes personal "identifiable" information and therefore must be securely administered in full compliance with all applicable federal and state privacy laws and regulations, including but not limited to, the Health Insurance Portability and Accountability Act (HIPAA).

### 3. Software (or geocoding application)

Software, or the geocoding application, can perform many functions. The basic function is to parse the source data into defined elements for better understanding and matching. In a physical address example this parsing includes the address number, street prefix, street name, the street suffix, street direction, street type (e.g. BLVD), City name, ZIP Code, ZIP+4, and State. Software then performs a probabilistic record linkage with a statistically valid form of fuzzy logic to score how well the source data can be matched to the Reference Data. This type of matching allows for reviews of "almost" matches, scoring thresholds, index tuning, best match and/or candidate matching.

Geocoding software can also perform a standardization process, whereby the source data is standardized along, for example, the US Postal Service standards. This process increases the match rate potential of the source data and provides for a common storage taxonomy for the source data. Geocoding software can return, depending on the solution architecture, an X, Y location, the standardized address, a score, a match sequence (e.g. what reference data it matched to), and ancillary data as required.

It is important to distinguish between software designed to geocode for the purpose of performing spatial analysis and software designed to geocode for the purpose of minimizing the expenses related to undeliverable and duplicate mail.

## Standard Requirements

The following standard is approved for State of California geocoding methods. This standard is a 'process' standard. The description below identifies the appropriate steps to follow to comply with the state standard for managing address data.

### *Step 1: Define Input or Source Data – Field Definitions*

Please note: Source data can contain errors in address formatting or transposed elements (e.g. numbers entered incorrectly or transposed, misspelled addresses, not real addresses and non-standardized address elements (e.g. Boulevard vs. BLVD). These errors should be corrected through a standardization process at the onset of geocoding (i.e., data cleansing).

It is the standard of the State of California for agencies to store a minimum of the following fields for descriptive location (e.g. address) data. It is acceptable to modify the field definitions to meet specific business needs (e.g. change field name and/or lengths). However, it is not acceptable to not carry one of the below fields along with the geocoded record.

| Field | Type | Description |
|---|---|---|
| First Address Field | Text | Street number, and name, intersections acceptable.<br>Example: 1325 J Street |
| Second Address Field | Text | Building, floor, mail stop, PO Box, suite, etc.<br>Example: Suite 1600 |
| City | Text | City Name<br>Example: Sacramento |
| State | Text | State Abbreviation<br>Example: CA |
| ZIP Code | Text | ZIP Code<br><br>Example: 95814 |
| ZIP 4 Extension (optional) | Text | ZIP Code + "-" and 4 digit extension if available<br><br>Example: 95814-1234. |

### *Step 2 – Standardization and Validation Process*

Once data is assembled in the proper format, it is the standard of the State of California for geocoding processes to first standardize and validate the accuracy of street address records according to current United States Postal Service (USPS) address data by means of a USPS Coding Accuracy Support

System (CASS) Certified (http://www.usps.gov/cass.htm) software product. CASS-certified software can be part of a single geocoding package, a stand alone software product and/or a web service.

***Step 3: Composite Geocoding Process***

A composite geocoding service will rely on multiple reference data, and in order of best quality attempt to match the source data in a cascading fashion. If no match is reached in the best quality reference data set, the service will turn to the next reference data set and attempt to perform the same function and repeat until a match is found which meets the business rules applied to the reference data. A composite geocoding service performs the same function as a single reference geocoding service plus the reference layer it matched to. In this environment record level analysis can then determine the rate for best quality versus lower quality matches in addition to finding some location for nearly every single source data record rather than just leaving them Unmatched. The State of California standard geocoding application is a composite service to ensure the most possible location matches.

***Step 4 – Keep Post Geocoded Address Fields***

It is the standard of the State of California for agencies to keep a minimum of the following fields from the results of the geocoding process. It is acceptable to modify the field definitions to meet specific business needs (e.g. change field name and/or lengths). However, it is not acceptable to delete any of the below fields along with the geocoded record. Typical industry geocoding software will return all of these fields.

| Field | Type | Width | Description |
|---|---|---|---|
| Geocoding Score | Integer | 3 | A score rating the quality of address match (geocode) to a reference data set. Business rules for minimum match score are the responsibility of the sponsoring agency for defining.<br><br>Example: 100 |
| Geocoding reference data | Text | 50 | A string indicating the source of the reference data to which the address was matched<br><br>Example: 1-TA_Points_ZIP_0708 |
| Standard Address | Text | 50 | The standardized and validated address captured for address data quality and management (output of Address 1).<br><br>Example: 2575 Sand Hill Rd |
| Standard City Name | Text | 30 | The standardized and validated city name captured for data quality and management (output of City).<br><br>Example: Davis |
| Longitude | Floating decimal | 8 | The x-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need.<br><br>Example: -120.554987 |
| Latitude | Floating decimal | 8 | The y-coordinate of the geocoded address in geographic projection (NAD83 – see projection standard). A minimum of 6 decimals must be carried in this field, depending on the business need.<br><br>Example: 37.491958 |
| X | Floating decimal | 3 | The x-coordinate of the geocoded address in standard California Albers Projection (NAD83 – see projection standard). A minimum of 3 decimal places must be carried in this field.<br><br>Example: -100125.1234 |
| Y | Floating decimal | 3 | The y-coordinate of the geocoded address in standard California Albers Projection (NAD83 – see projection standard). A minimum of 3 decimal places must be carried in this field.<br><br>Example: 43678.1234 |

***Step 5 – Final Database Management***

Agencies are able to maintain these fields in data models suiting their appropriate needs, as long as a minimum number of the above fields are maintained at the record level.

## Definitions

Below are definitions pertinent to the geocoding processes that are included in the SIMM 58C, Enterprise Architecture Glossary:

**Geocode** – A standardized representation for a location given a textual description of the location like an address, ZIP Code, or place name. The standardized representation is typically an X, Y coordinate and/or a latitude and longitude. Generally speaking, geocode refers only to a street address text description of place.

**Composite Service** – a service used to provide a geocode address whereby multiple layers of address data are used in a hierarchy to achieve maximum accuracy.

**Coding Accuracy Support System (CASS)** – The CASS is the United States Postal Service Coding Accuracy Support System. The CASS enables the USPS to evaluate the accuracy of address-matching software programs and provide grades for vendors of software. In addition, the vendors then have an ability to change and modify their software to increase the accuracy of address-matching functions. Many currently available software packages meet the CASS standard.

## Authorities

As described in Government Code Section 11545, the OCIO has broad responsibility and authority to guide the application of Information Technology in California State government. This includes establishing and enforcing state IT strategic plans, policies, and standards.

Geographic Information Systems (GIS) have a significant IT component, and thus fall with the jurisdiction of the OCIO.

## Implementation

This EA Standard applies to all new data system development for IT projects approved beginning in January 2010 that are initially funded in the Budget Act of 2010.

For systems that are already in place, state agencies should review the EA Standard, and incorporate implementation or retrofit plans into their Agency Information Management Strategy.

Exceptions to this EA Standard may be submitted to the OCIO by following the "OCIO EA Compliance Component Instructions" found in the SIMM 58A, Enterprise Architecture Developers Guide.

Data stored in individual desktop productivity tools, such as spreadsheets, is not subject to this EA standard. However, agencies interested in geocoding such data for mapping purposes are encouraged to follow the EA Standard and associated EA Practice.